STA 313: Topics in Statistics

Al Nosedal.
University of Toronto.

Fall 2015

"essentially, all models are wrong, but some are useful"

George E. P. Box

(one of the great statistical minds of the 20th century).

What is R?

The R system for statistical computing is an environment for data analysis and graphics.
The main source of information about the R system is the world wide web with the official home page of the R project being
http://www.R-project.org
All resources are available from this page: the R system itself, a collection of add-on packages, manuals, documentation and more.

Installing R

The R system for statistical computing consists of two major parts: the base system and a collection of user contributed add-on packages. A package is a collection of functions, examples and documentation. Both the base system and packages are distributed via the Comprehensive R Archive Network (CRAN) accessible under
http://CRAN.R-project.org

The Base System and the First Steps

The base system is available in source form and in precompiled form for various Unix systems, Windows platforms and Mac OS X. For us, it will be sufficient to download the precompiled binary distribution and install it locally. Just go to
http://CRAN.R-project.org
download the corresponding file (Download R for Linux or Download R for (Mac) OS X or Download R for Windows), execute it locally and follow the instructions given by the installer.

Help

The help system is a collection of manual pages describing each user-visible function and data set that comes with R. A manual page is shown in a pager or web browser when the name of the function we would like to get help for is supplied to the help function

```
help("mean")
```

Elementary commands

R code

```
42+8;
8-2;
8*2;
8/2;
# lines preceded by # are comments and ignored by R;
```

Vectors

Vectors may be created in several ways, of which the most common is via the c function which combines all values from all arguments to the function.

R code

```
x=c(1,2,3,4);
# this is an assignment;
# indicated by the operator = ;
x;
y=c(6,7,8);
c(y,x,y);
```

Arithmetic operators applied to vectors

R code

```
x*x;
# you should get:  1 4 9 16;
x/x;
# you should get:  1 1 1 1;
# The number of elements in a vector is
# extracted by the length function;

length(x);
# you should get 4;
```

Matrices

R code

```
x=c(1,2,3,4,5,6);
X=c(10,11,12,100,-5,-6);
M=matrix(c(x,X),ncol=2);
# the vectors x and X are
# combined using the c function;
# and then converted to a 2-column
# matrix by the matrix function;
M;
matrix(c(1,2,3,4,5,1,2,3,4),ncol=3);
```

Arithmetic operations applied to Matrices

R code

```
M*M;

M/M;
```

Matrix multiplication

Matrix multiplication can be achieved using the %*% operator.

R code

```
M1=matrix(c(1,2,1,2),ncol=2);
# M1 will be a 2 x 2 matrix;
M2=matrix(c(3,4,5,6,7,8),ncol=3);
# M2 will be a 2 x 3 matrix;
M1%*%M2;
# the transpose of M2 multiplied
# by M1 can be obtained using
# the t function;
t(M2)%*%M1;
```

Extracting elements of a Matrix

R code

```
Mrow1=M[1, ];
Mrow1;
M[ ,2];
M[1,2];
```

"Subsetting" of vectors and matrices

R code

```
M3=matrix(1:10,ncol=5);
# the command 1:10 generates a vector
# containing the elements 1 to 10;
M3;
M3sub=M3[ ,c(1,3,5)];
# selects columns 1,3, and 5;
M3sub;
```

"Combining" matrices

R code

```
X1=matrix(1:10,ncol=2);
# X1 is 5 x 2;
Y1=matrix(1:20,ncol=4);
# Y1 is 5 x 4;
Z1=cbind(X1,Y1);
Z1;
rbind(X1,Y1[ ,1:2] );
```

Example

The following table lists the top 10 countries and amounts of oil (millions of barrels annually) they exported to the United States in 2010.

Country	Oil Imports (millions of barrels annually)
Algeria	119
Angola	139
Canada	720
Colombia	124
Iraq	151
Kuwait	71
Mexico	416
Nigeria	360
Saudi Arabia	394
Venezuela	333

Example (cont.)

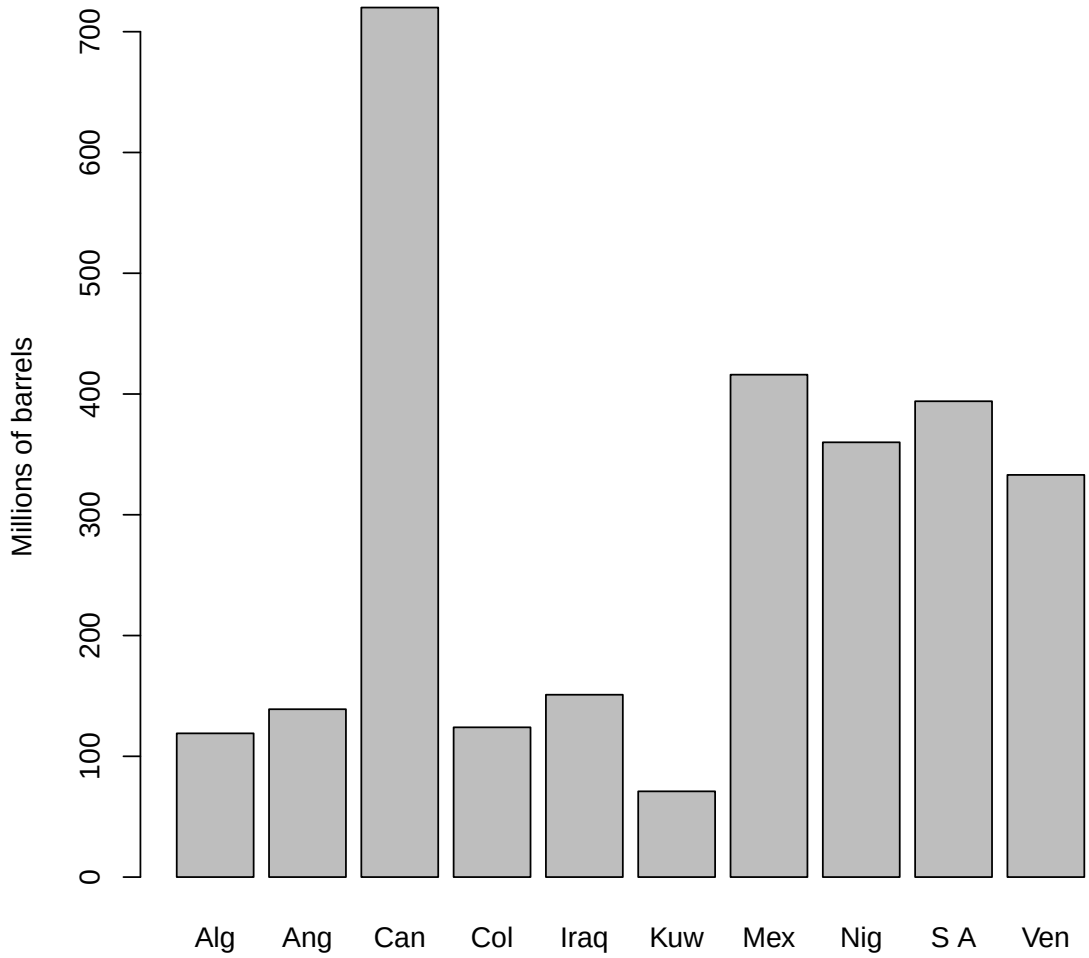
a. Draw a bar chart.
b. Draw a pie chart.

Solution (Bar chart)

R code

```
# Step 1.  Entering data;
barrels=c(119,139,720,124,151,71,416,360,394,333);
country=c("Alg","Ang","Can","Col","Iraq","Kuw","Mex",
"Nig","S A","Ven");
# Step 2.  Making bar chart;
barplot(barrels,names.arg=country,ylab="Millions of
barrels");
```

Bar chart

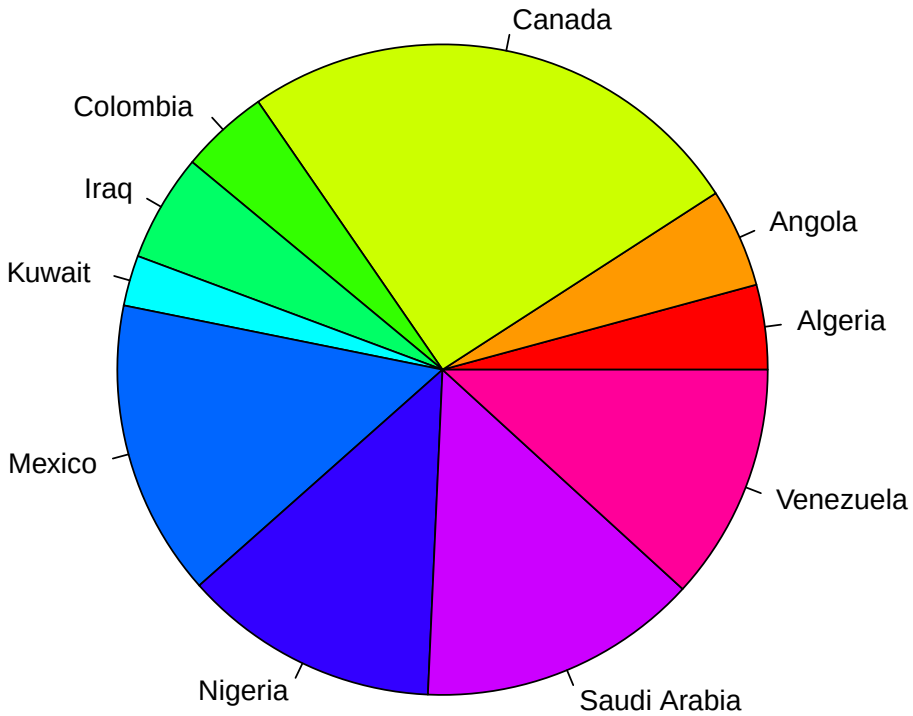

Millions of barrels
0
100
200
300
400
500
600
700
Alg
Ang
Can
Col
Iraq
Kuw
Mex
Nig
S A
Ven

Solution (Pie chart)

R code

```
# Step 1.  Entering data;
barrels=c(119,139,720,124,151,71,416,360,394,333);
country=c("Alg","Ang","Can","Col","Iraq","Kuw","Mex",
"Nig","S A","Ven");
# Step 2.  Making pie chart;
pie(barrels,country,col=rainbow(10));
```

Pie chart


Canada
Colombia
Iraq
Kuwait
Angola
Algeria
Mexico
Venezuela
Nigeria
Saudi Arabia

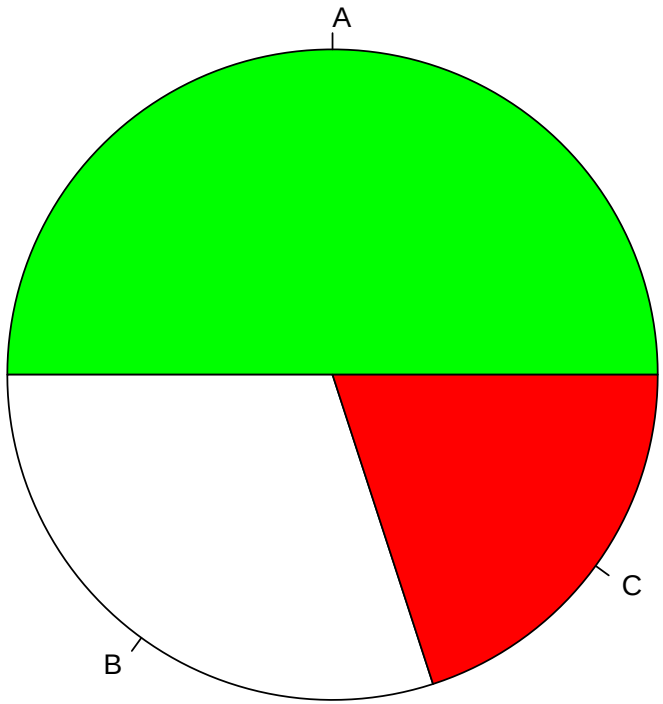
Exercise

The response to a question has three alternatives: A, B, and C. A sample of 120 responses provides 60 A, 24 B, and 36 C.
a)Show the frequency, relative frequency and percent frequency distributions.
b) Construct a pie chart.
c) Construct a bar graph.

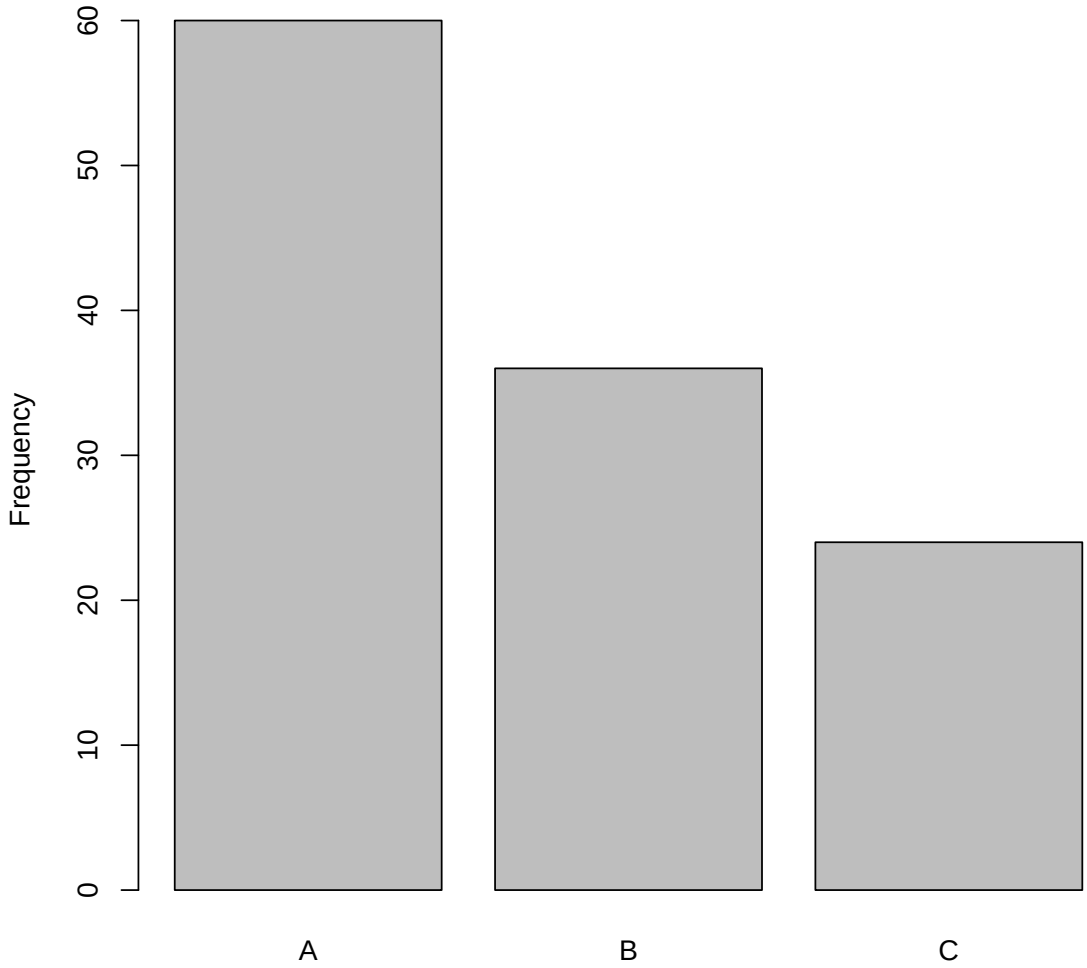
Solution

Class	Frequency	Relative Freq.	Percent Freq.
A	60	60/120	0.50
B	24	24/120	0.20
C	36	36/120	0.30

Solution (pie chart)



Solution (bar chart)



Exercise. Never on Sunday?

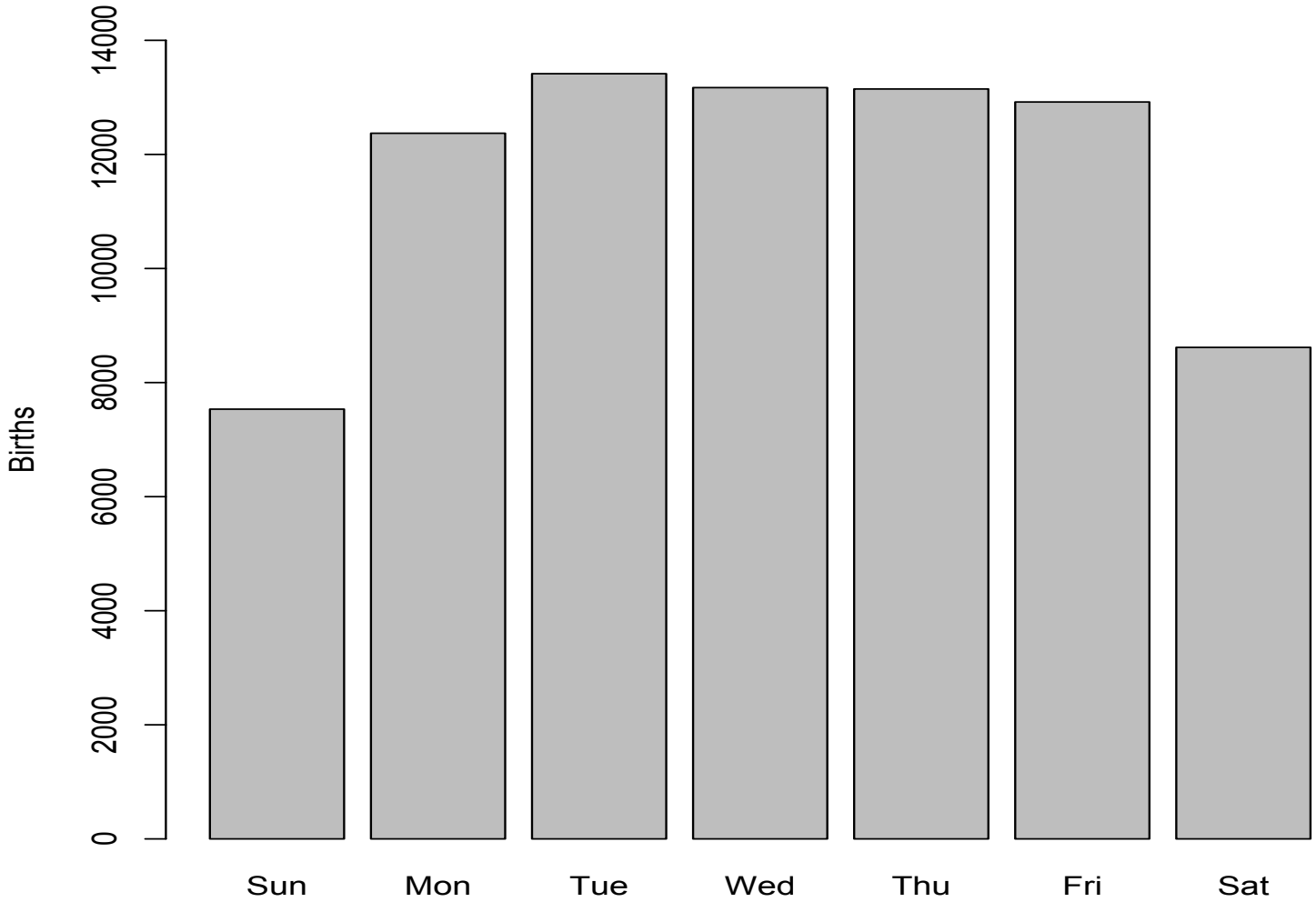
Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2008:

Day	Births
Sunday	7,534
Monday	12,371
Tuesday	13,415
Wednesday	13,171
Thursday	13,147
Friday	12,919
Saturday	8,617

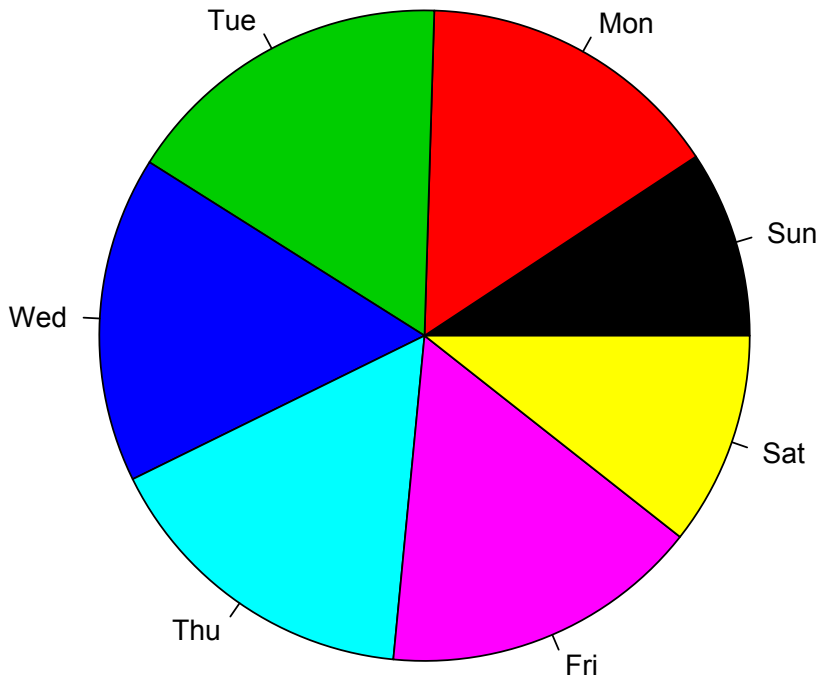
Exercise. Never on Sunday? (cont.)

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

Exercise. Never on Sunday? Bar chart.



Exercise. Never on Sunday? Pie chart.



Exercise. Never on Sunday?

Solution.
It would be correct to make a pie chart but a pie chart would make it more difficult to distinguish between the weekend days and the weekdays. Some births are scheduled (e.g., induced labor), and probably most are scheduled for weekdays.

Exercise. What color is your car?

The most popular colors for cars and light trucks vary by region and over time. In North America white remains the top color choice, with black the top choice in Europe and silver the top choice in South America. Here is the distribution of the top colors for vehicles sold globally in 2010.

Color	Popularity (%)
Silver	26
Black	24
White	16
Gray	16
Red	6
Blue	5
Beige, brown	3
Other colors	

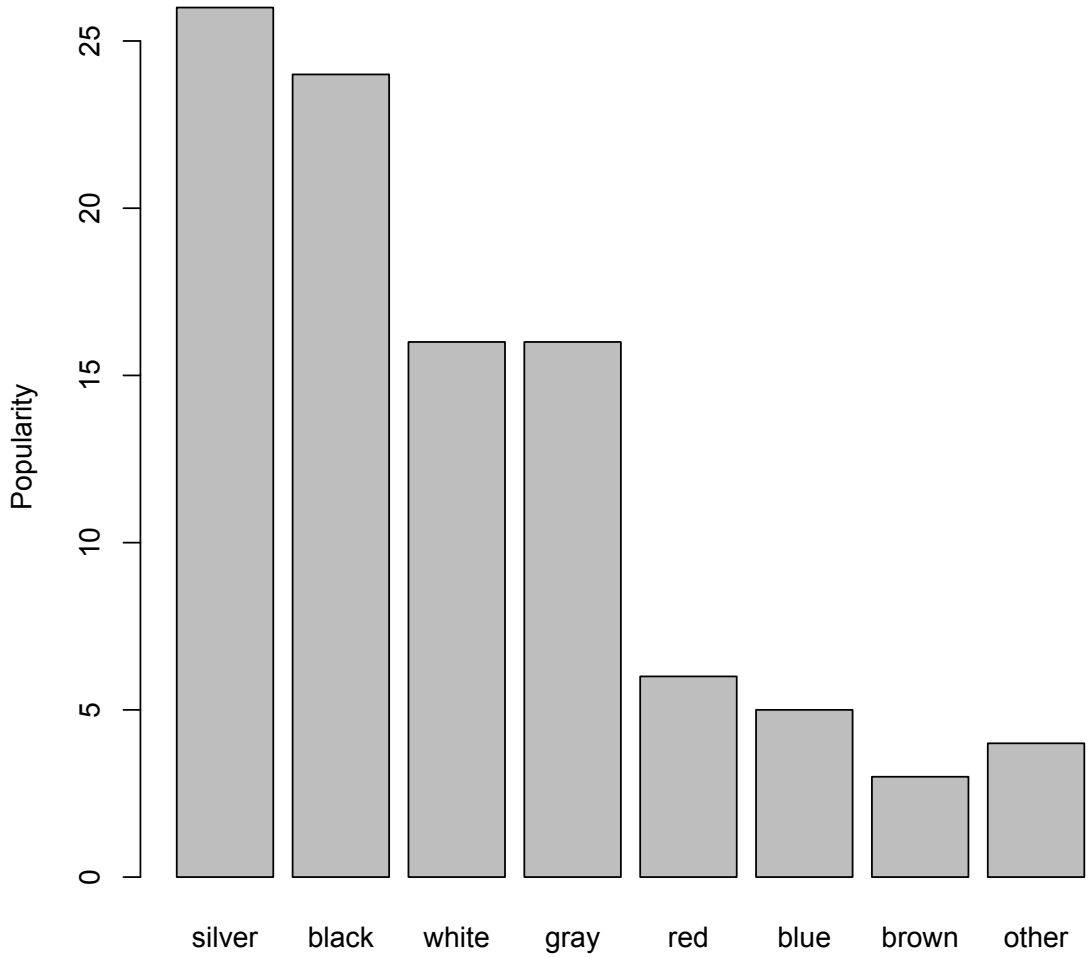
What color is your car? (cont.)

a) Fill in the percent of vehicles that are in other colors.
b) Make a graph to display the distribution of color popularity.

Solution

a) $\text{Other} = 100 - (26 + 24 + 16 + 16 + 6 + 5 + 3) = 4.$

Graph


Popularity
0
5
10
15
20
25
silver
black
white
gray
red
blue
brown
other

Newspaper Readership Survey

A major North American city has four competing newspapers: the *Globe and Mail (G & M)*, *Post*, *Star*, and *Sun*. To help design advertising campaigns, the advertising managers of the newspapers need to know which segments of the newspaper market are reading their papers. A survey was conducted to analyze the relationship between newspapers read and occupation. A sample of newspaper readers was asked to report which newspaper they read - *Globe and Mail (1), Post (2), Star (3), Sun (4)* - and indicate whether they were blue-collar workers (1), white-collar workers (2), or professionals (3).

Newspaper Readership Survey

Some of the data are listed here.

Reader	Occupation	Newspaper
1	2	2
2	1	4
3	2	1
⋮	⋮	⋮
352	3	2
353	1	3
354	2	3

Determine whether the two nominal variables are related.

Solution

By counting the number of times each of the 12 combinations occurs, we produced the following Table.

		Newspaper			
Occupation	**G & M**	**Post**	**Star**	**Sun**	**Total**
Blue Collar	27	18	38	37	120
White Collar	29	43	21	15	108
Professional	33	51	22	20	126
Total	89	112	81	72	354

Solution

Table of Row Relative Frequencies for our example.

		Newspaper			
Occupation	**G & M**	**Post**	**Star**	**Sun**	**Total**
Blue Collar	0.23	0.15	0.32	0.31	1
White Collar	0.27	0.40	0.19	0.14	1
Professional	0.26	0.40	0.17	0.16	1
Total	0.25	0.32	0.23	0.20	1

Solution

R code

```
# Step 1.  Entering data;
news.tab=matrix(c(0.23,0.27,0.26,0.15,0.40,0.40,
0.32,0.19,0.17,0.31,0.14,0.16),nrow=3,ncol=4);
news.tab;
```

Solution

R code

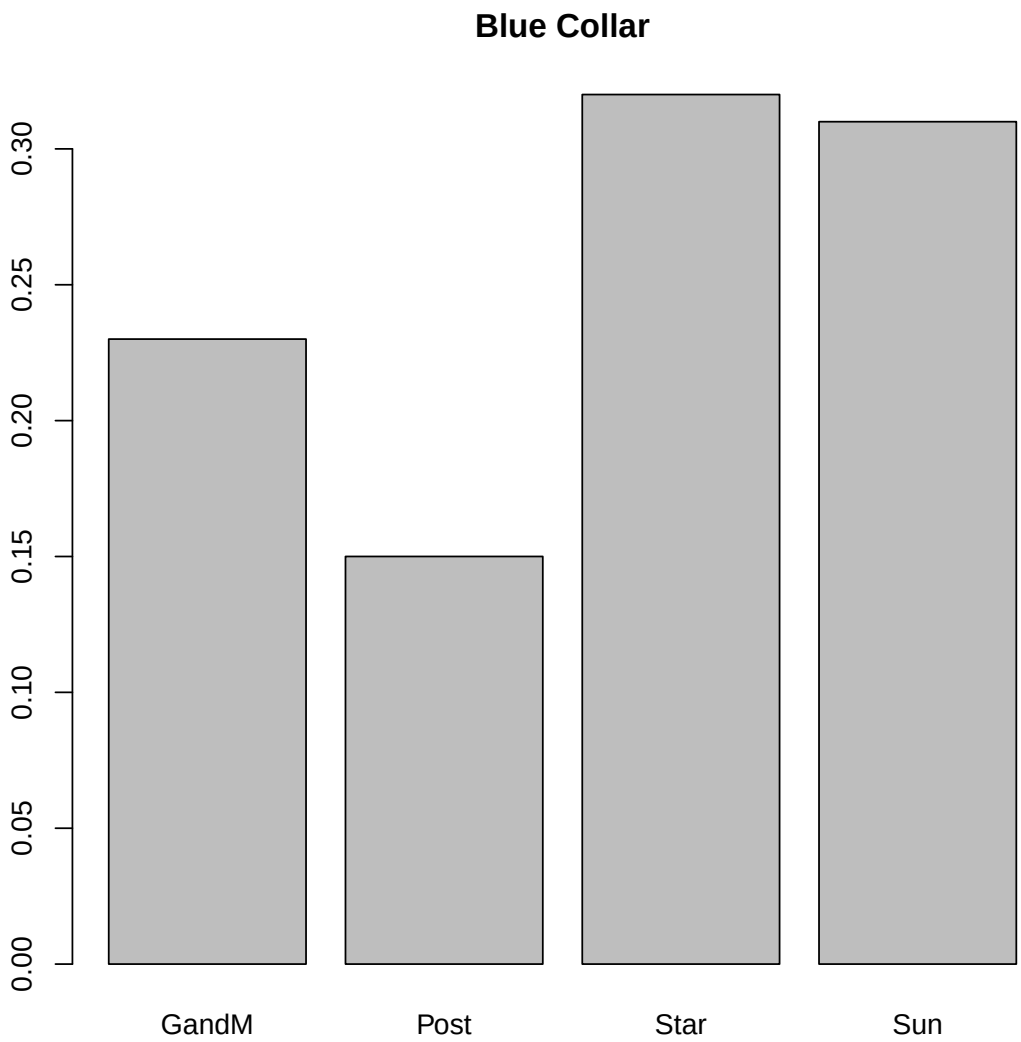
```
# Giving names to columns and rows;
colnames(news.tab)=c("GandM","Post","Star","Sun");
rownames(news.tab)=c("Blue Collar",
"White Collar", "Professional");
news.tab;
```

Solution

R code

```
# Step 2.  Bar chart for Blue Collar;
barplot(news.tab[1, ]);
title("Blue Collar");
```

Solution

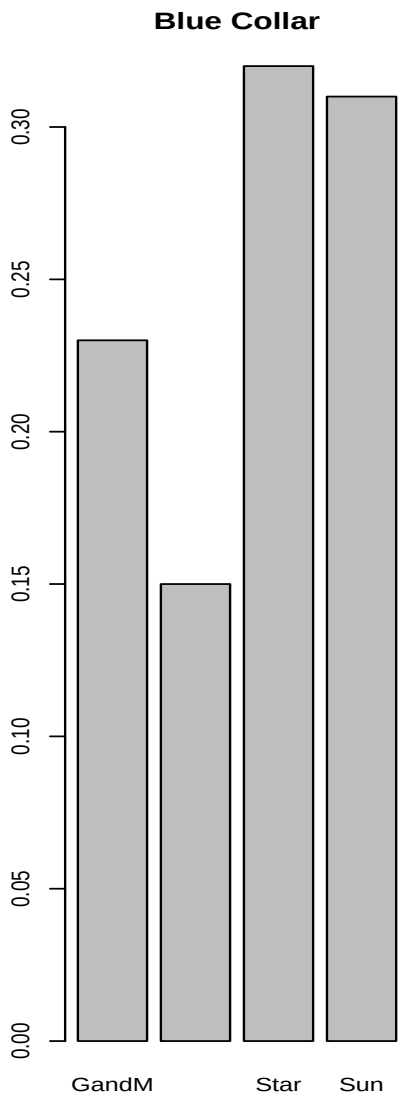


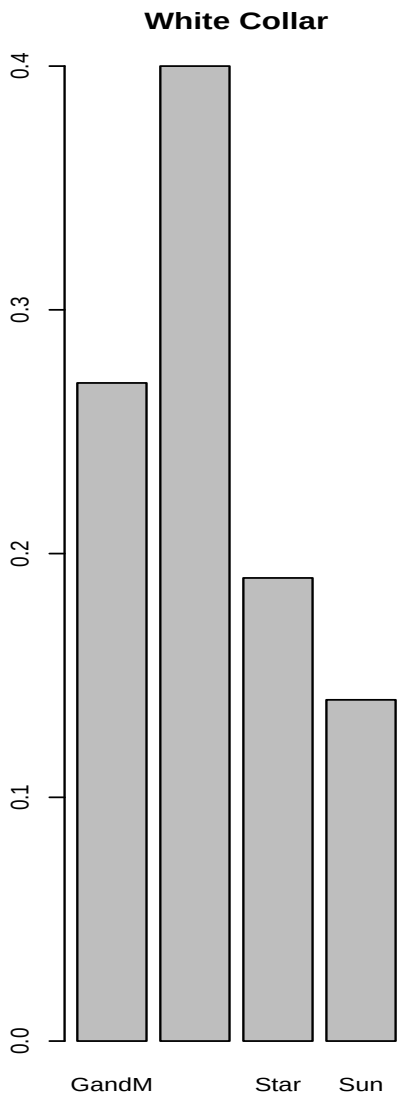
Solution (Side-by-Side bar charts)

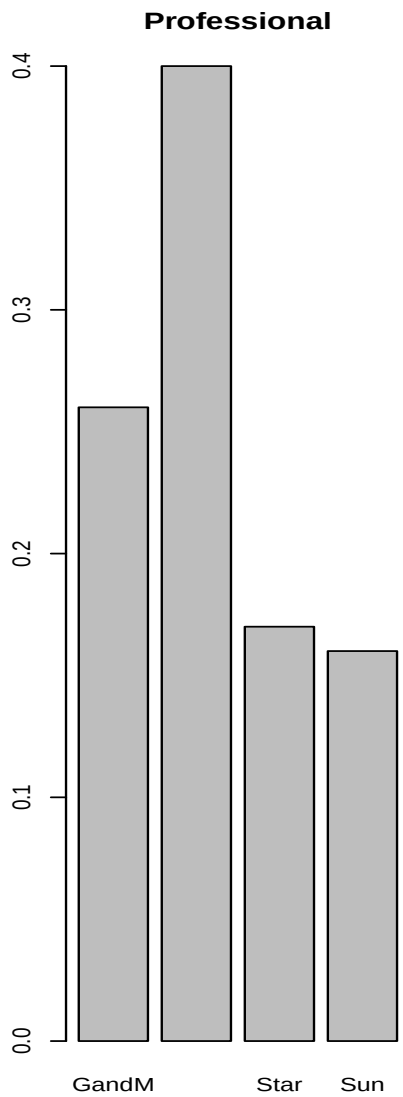
R code

```
par(mfrow=c(1,3));
barplot(news.tab[1, ])
title("Blue Collar");
barplot(news.tab[2, ])
title("White Collar");
barplot(news.tab[3, ])
title("Professional");
```

Solution


Blue Collar
0.00
0.05
0.10
0.15
0.20
0.25
0.30
GandM
Star
Sun


White Collar
0.0
0.1
0.2
0.3
0.4
GandM
Star
Sun


Professional
0.0
0.1
0.2
0.3
0.4
GandM
Star
Sun

Reading data from txt files

Now, we will learn how we can create tables from our data and calculate relative frequencies.

STEP 0. From your Desktop, create a new folder called: STA218

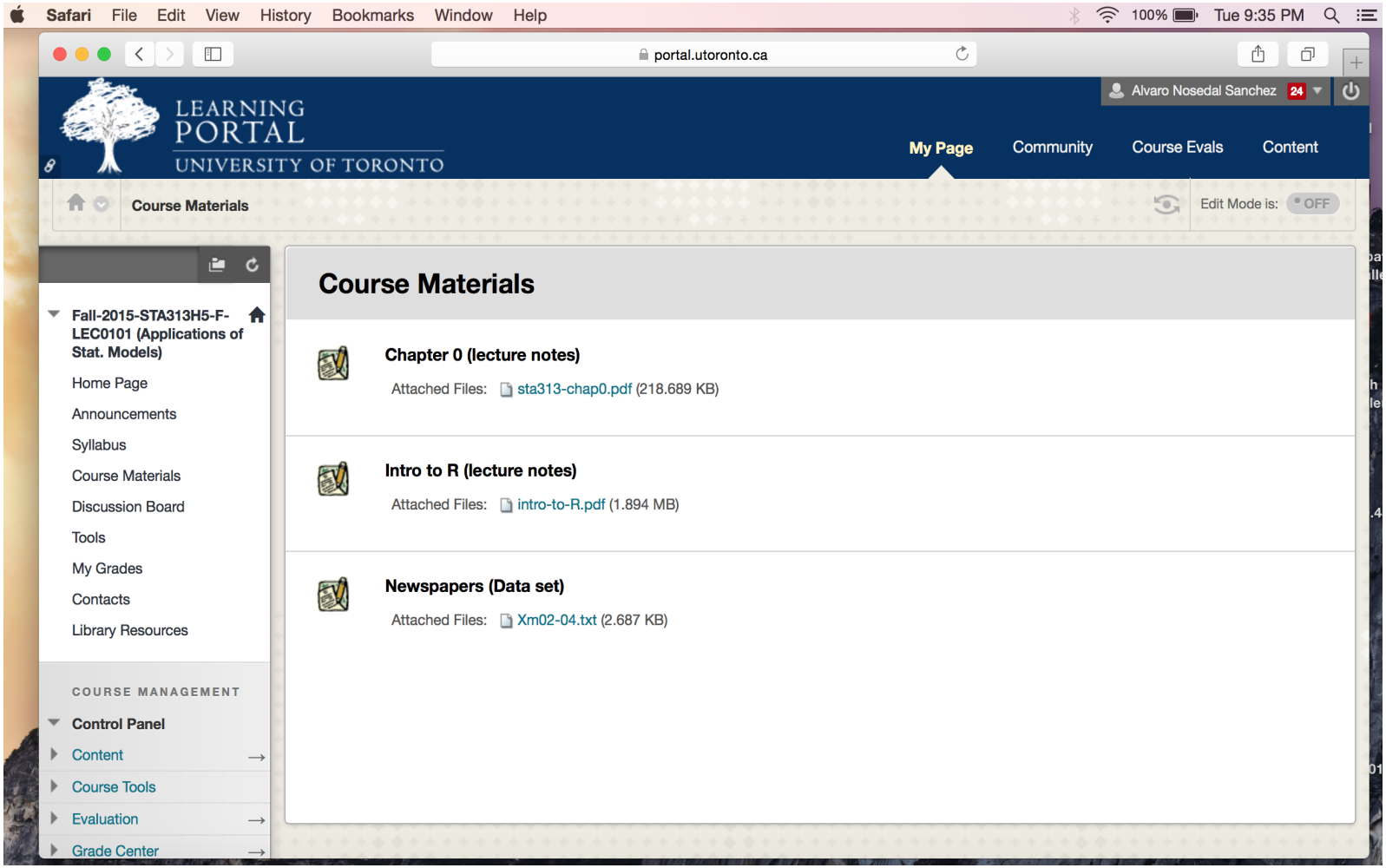
Reading data from txt files

STEP 1. Go to Portal
www.portal.utoronto.ca
STEP 2. Login.
STEP 3. Go to Fall-2015-STA313: Applications of Stat. Models

Reading data from txt files

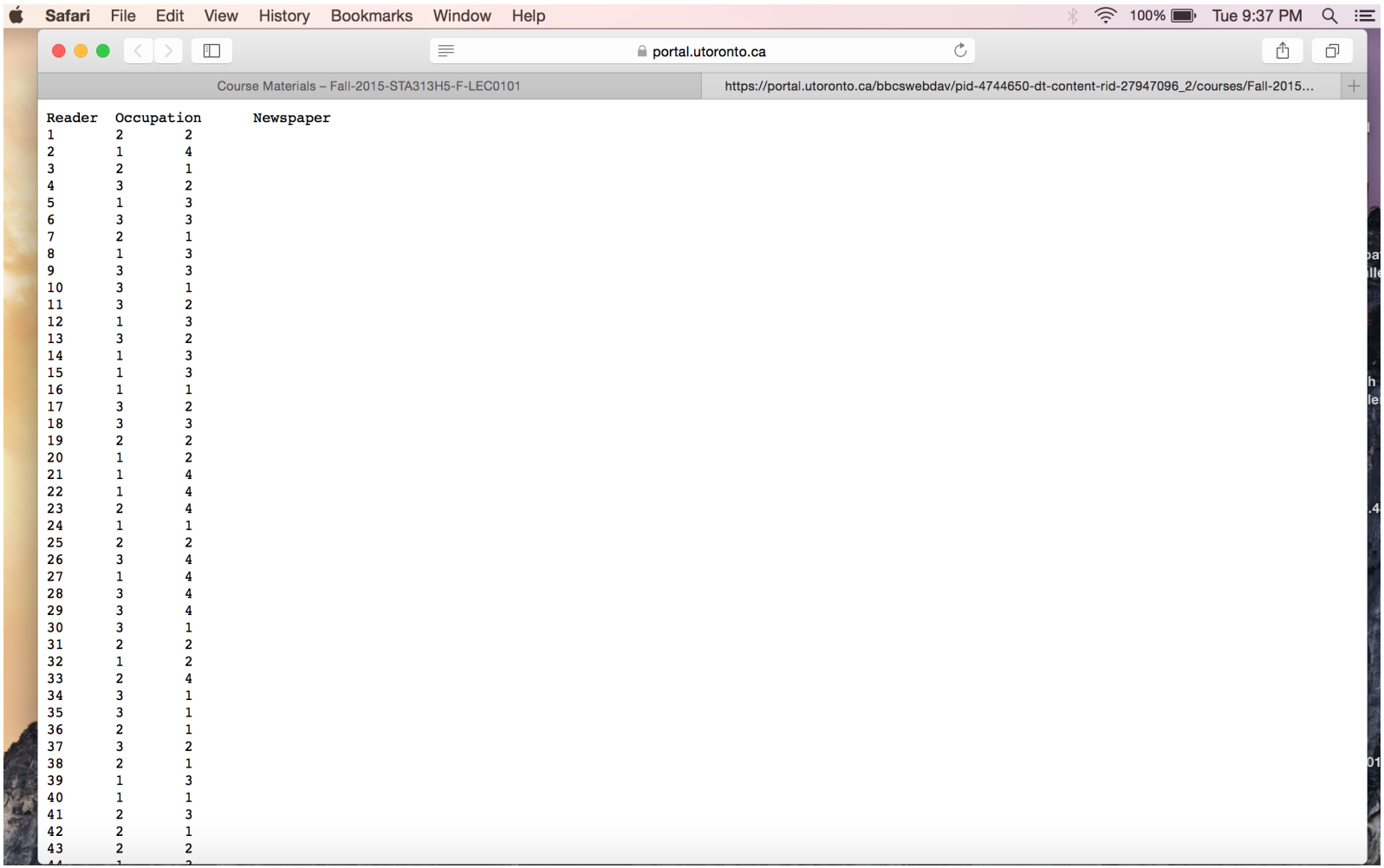
STEP 4. Go to Course Materials.

STEP 5. Find Newspapers (Data set).



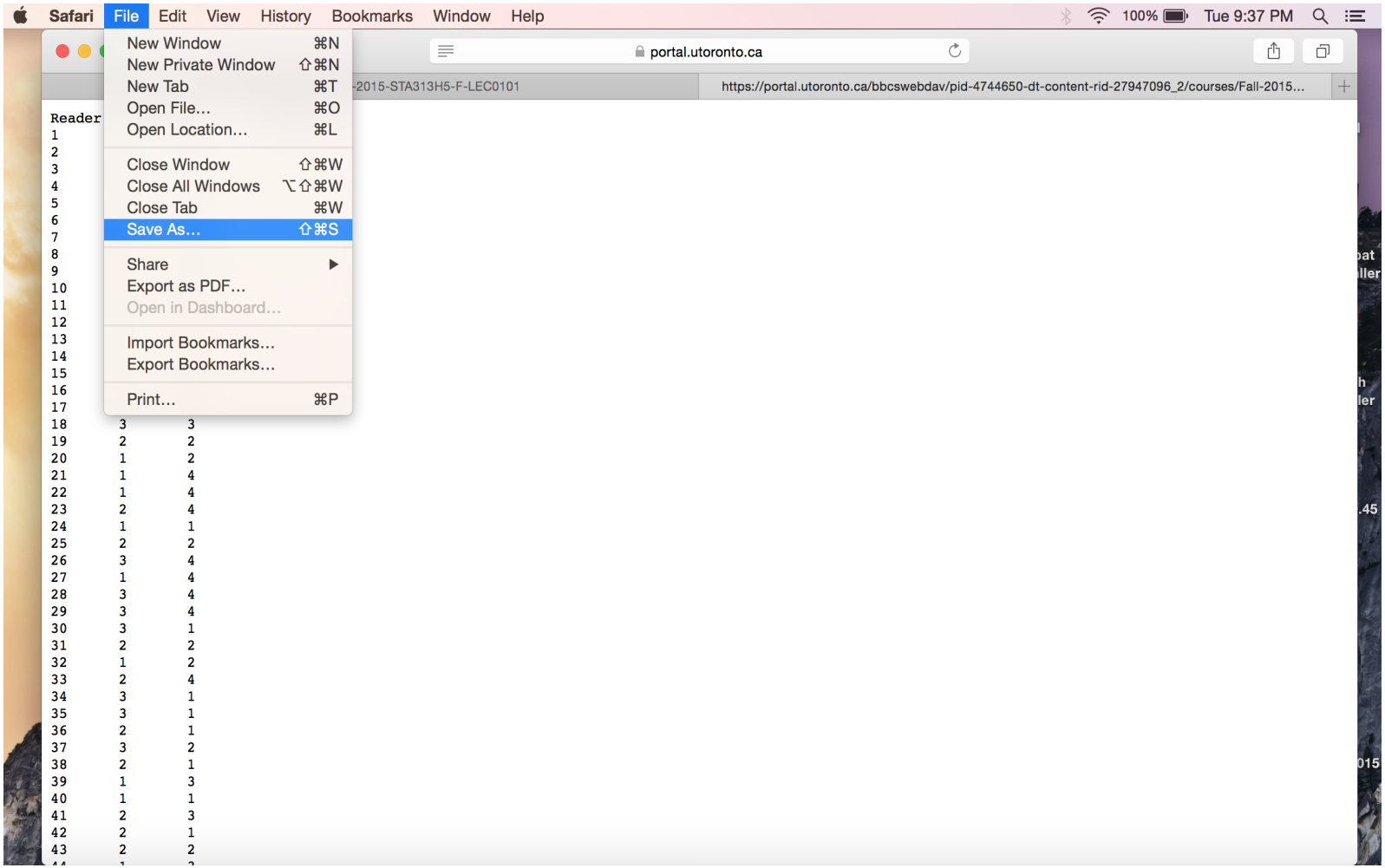
Reading data from txt files

STEP 6. Double-click on Xm02-04.txt (you should see something like this)



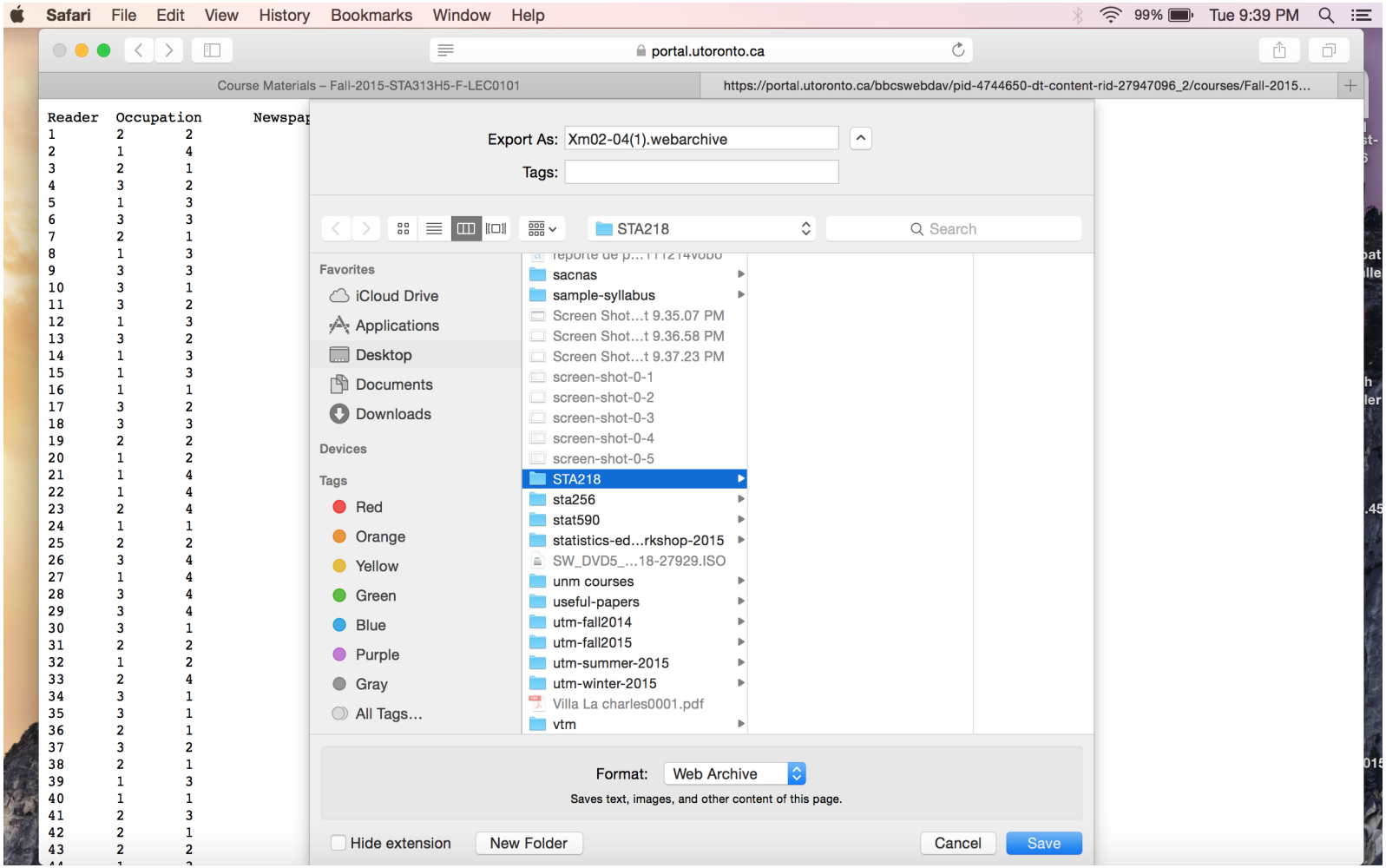
Reading data from txt files

STEP 7. Go to File. Please, select Save As...



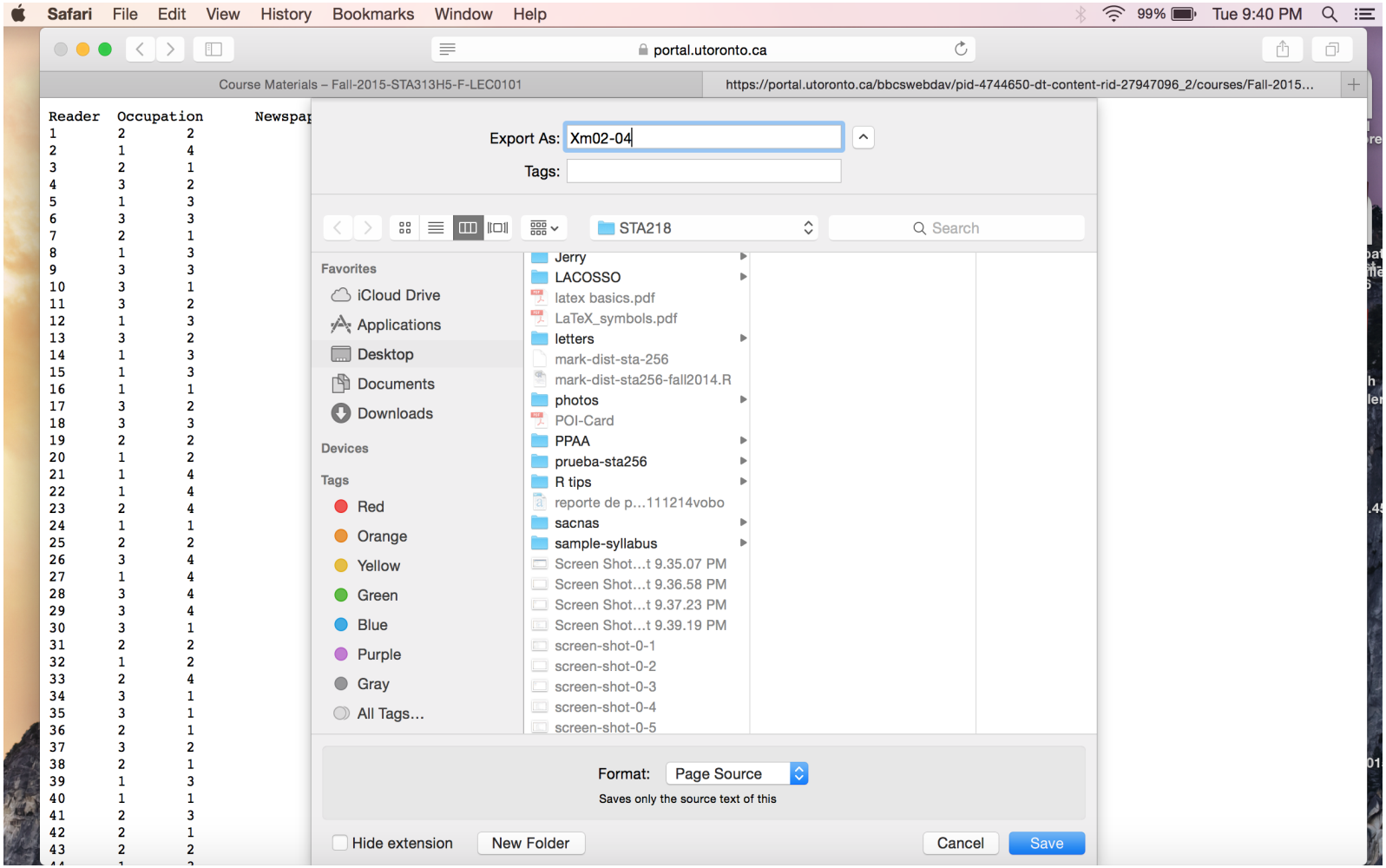
Reading data from txt files

STEP 8. Go to Format. Please, change Web Archive for Page Source



Reading data from txt files

STEP 9. Go to Format, change Web Archive for Page Source, and save Xm02-04.txt in STA218.

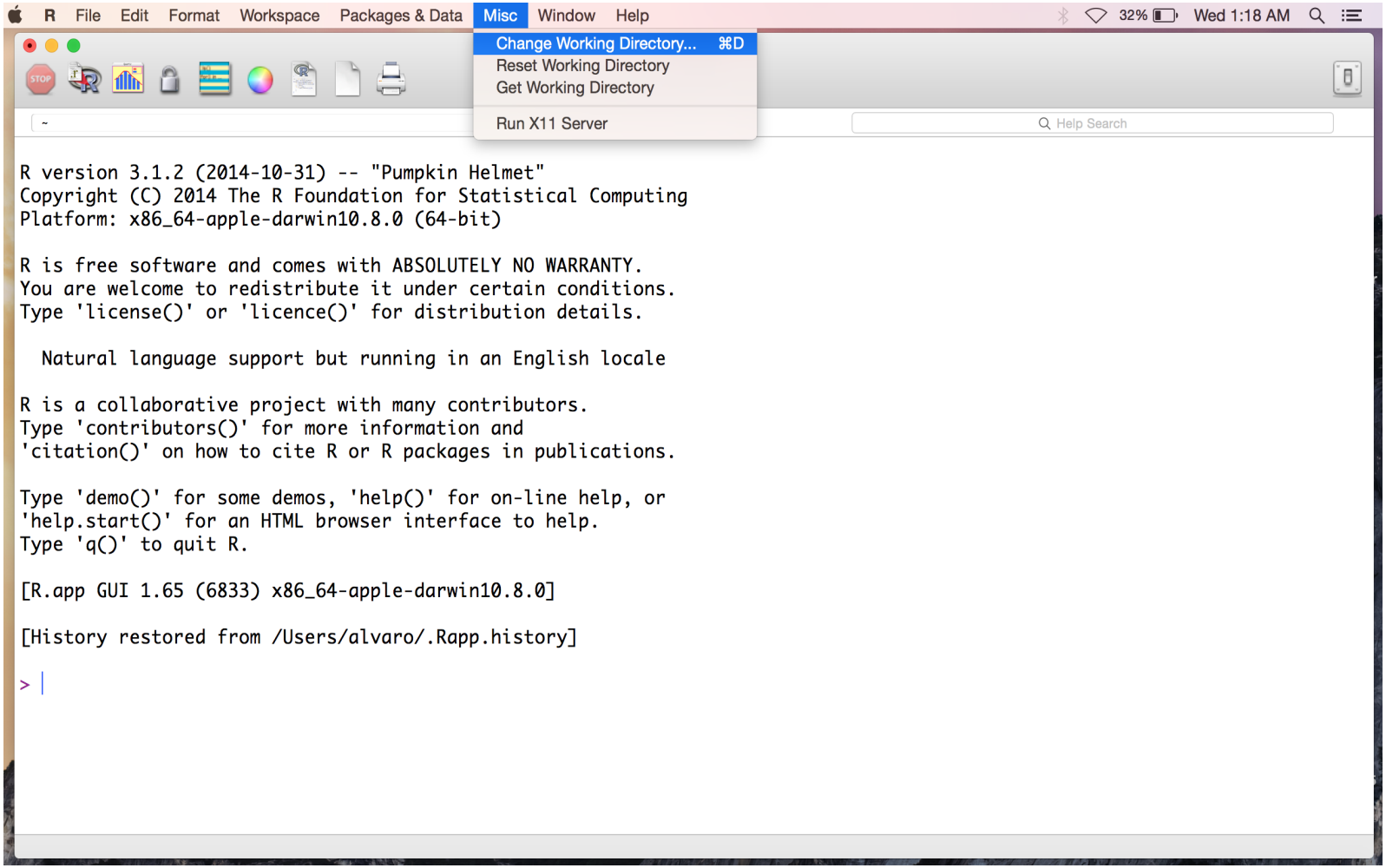


Reading data from txt files

STEP 10. Launch R

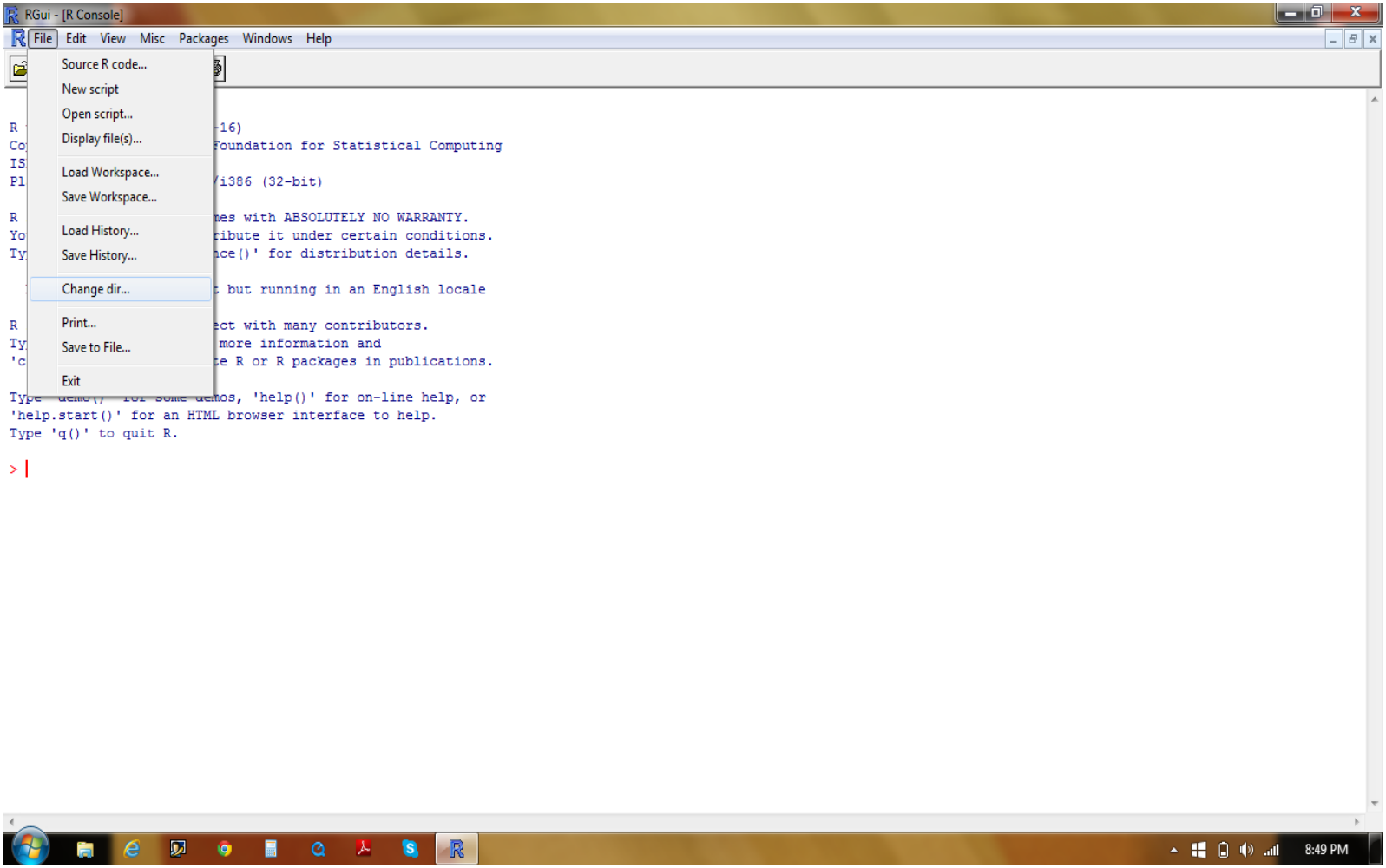
Reading data from txt files

STEP 11 (Mac users). Click on Misc (see screenshot) and select Change Working Directory ...



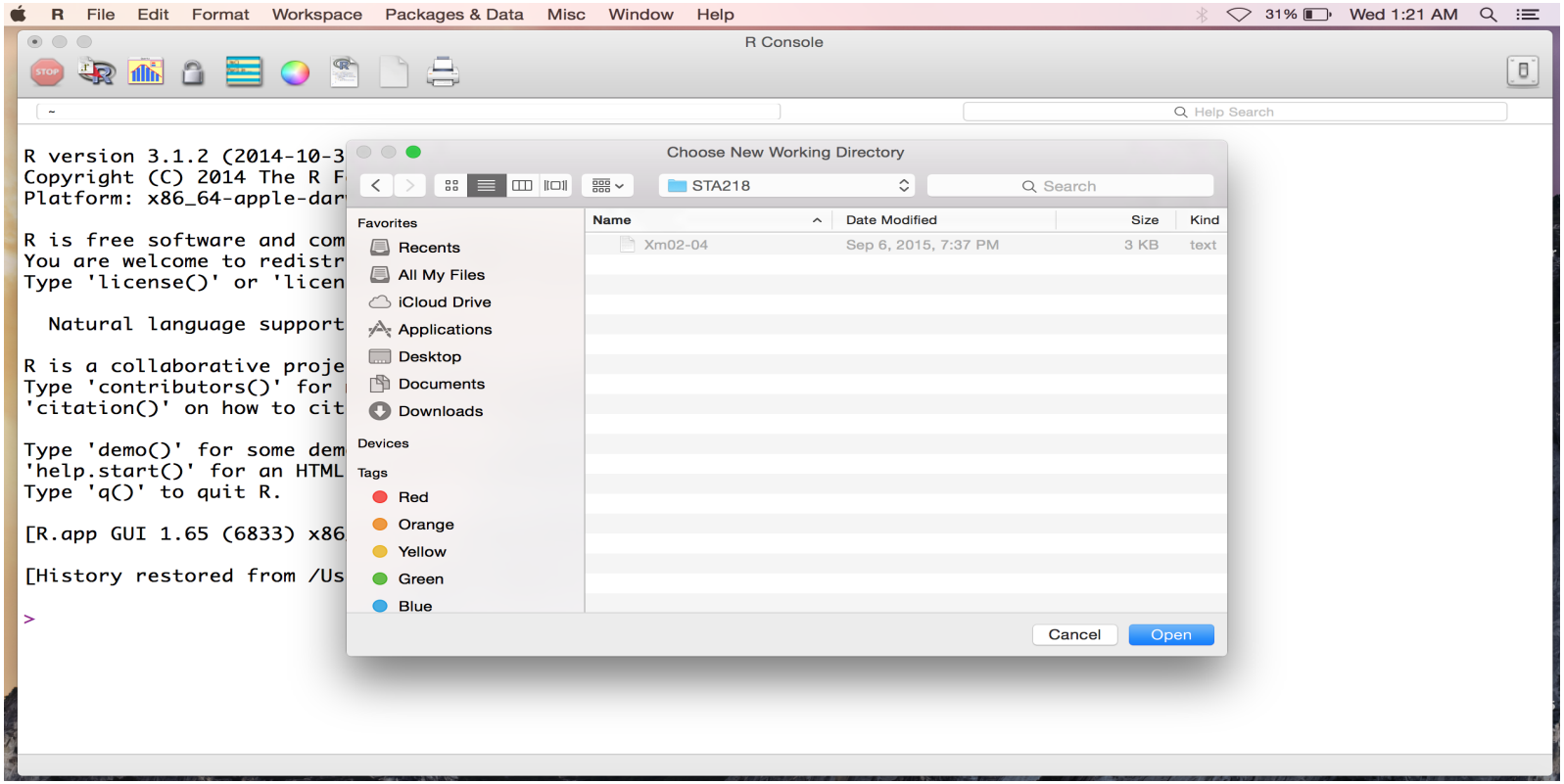
Reading data from txt files

STEP 11 (PC users). Click on File (see screenshot) and select Change dir...



Reading data from txt files

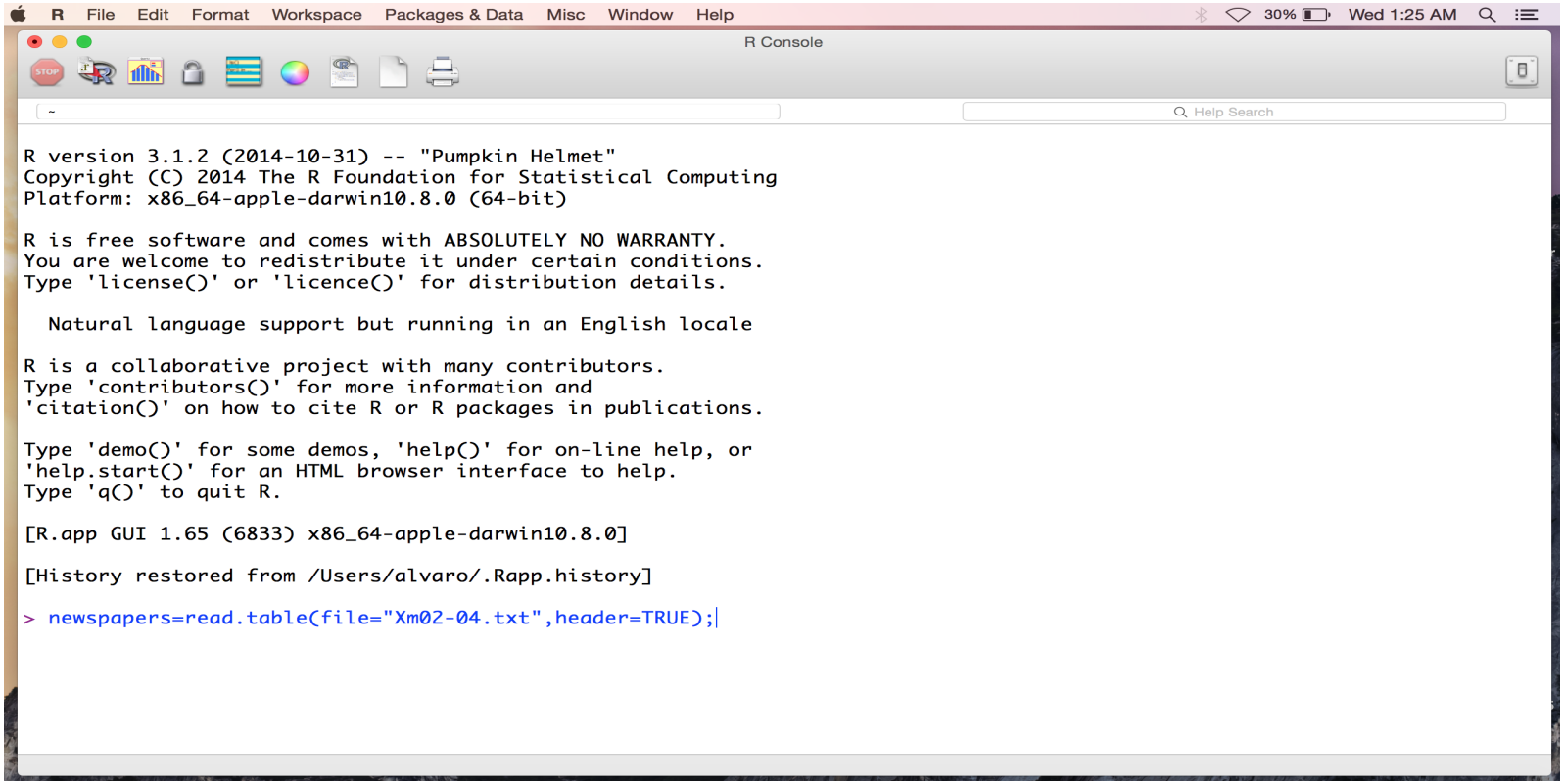
STEP 12. Find STA218. Then click on Open (see screenshot)



Reading data from txt files

STEP 13. Type the following in R Console:

```
newspapers=read.table(file="Xm02-04.txt",header=TRUE);
```



Press ENTER and . . . You are done!! (Reading the file)

Creating table of frequencies

R code

```
# Step 1.  "Reading" txt files;
newspapers=read.table(file="Xm02-04.txt",header=TRUE);
# Step 2.  Making table of frequencies;
xtabs(~ Occupation + Newspaper, data = newspapers);
```

Creating table of relative frequencies

R code

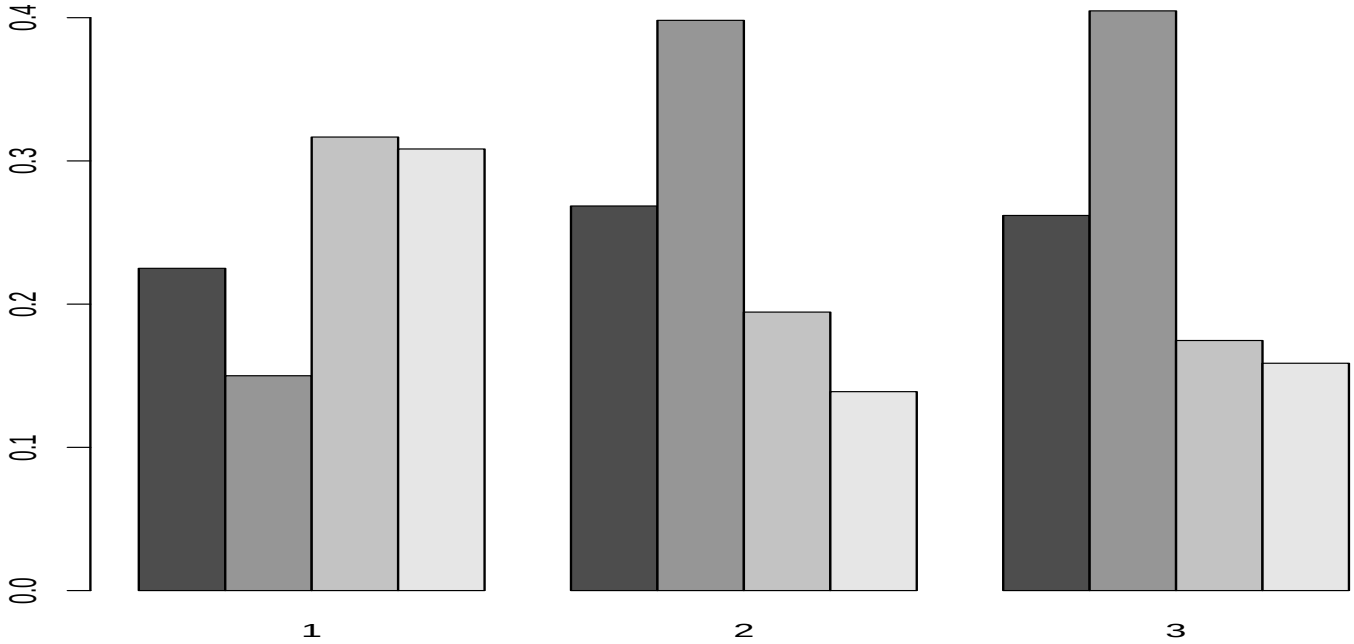
```
# Step 3.  Making table of relative frequencies;
freq.tab=xtabs(~ Occupation + Newspaper, data =
newspapers);
rel.freq.tab=prop.table(freq.tab,1);
rel.freq.tab;
```

Making bar charts

R code

```
# Step 4.  Graphing table of row relative
# frequencies;
barplot(t(rel.freq.tab),beside=T);
```

Bar charts



Built-in distributions in R

Four fundamental items can be calculated for a statistical distribution:

- Density or point probability
- Cumulative distribution function
- Quantiles
- Pseudo-random numbers

For all distributions implemented in R, there is a function for each of the four items listed above.

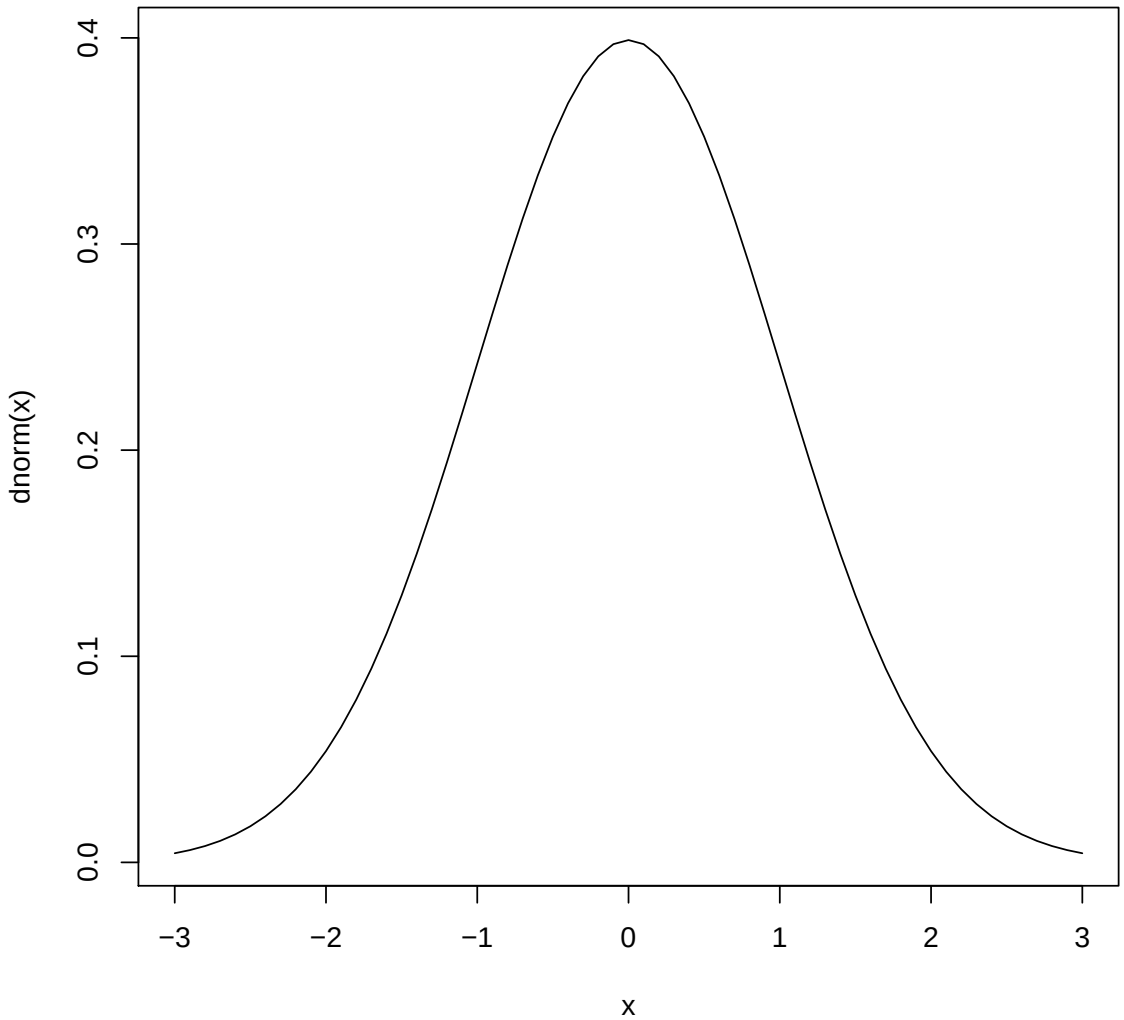
Probability density function (pdf)

If you want to draw the well-known bell curve of the Normal distribution, then it can be done like this:

R code

```
x=seq(-3, 3, 0.1);
plot(x, dnorm(x), type = "l");
# type = "l" causes the function to
# draw lines;
```

Normal pdf



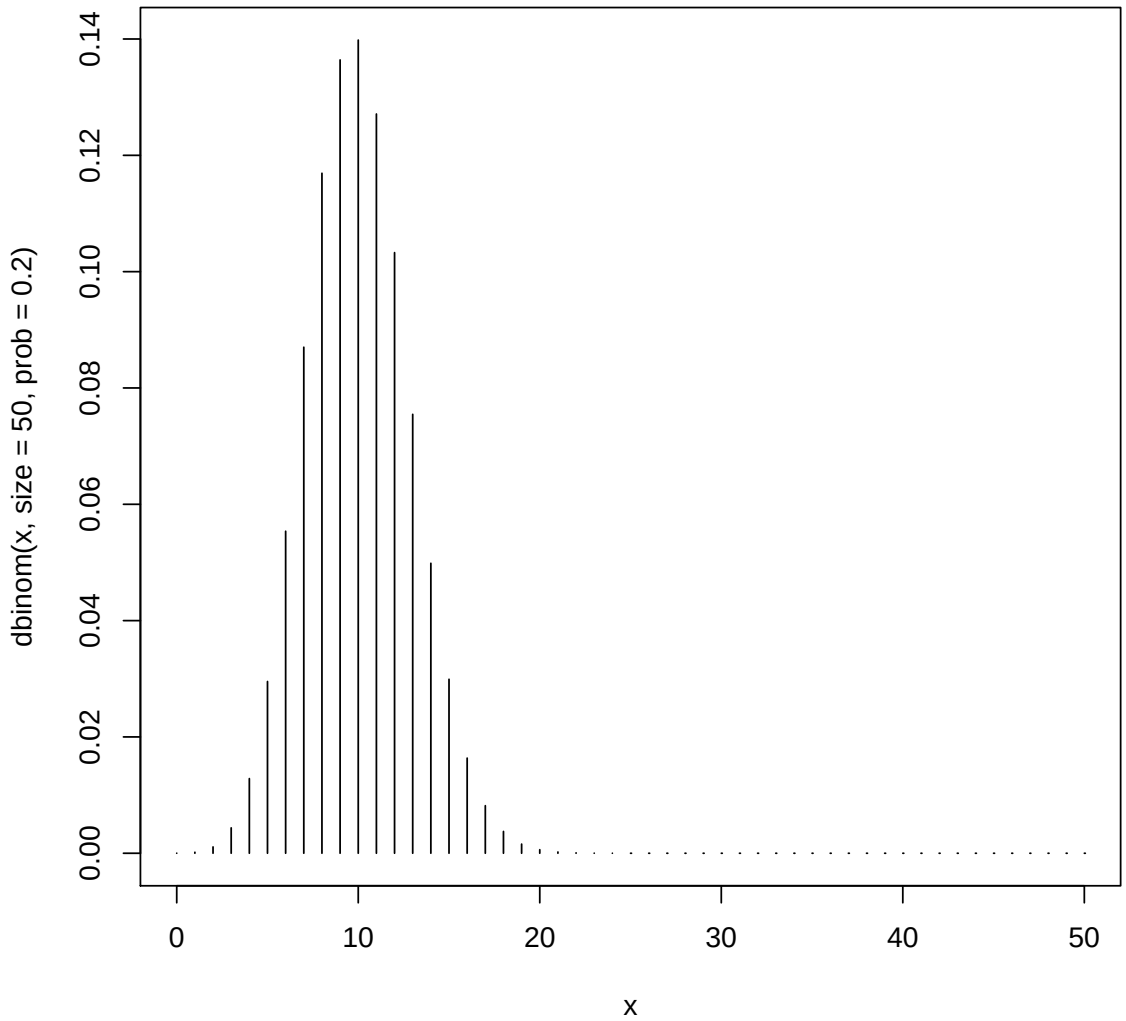
Probability mass function (pmf)

For discrete distributions it is preferable to draw a pin diagram, here for the Binomial distribution with $n = 50$ and $p = 0.20$

R code

```
x=seq(0,50,1);
plot(x, dbinom(x, size = 50, prob = 0.20), type =
"h");
# type = "h" causes the function to
# draw pins;
```

Binomial pmf



Cumulative distribution function (cdf)

The level of cholesterol in the blood is important because high cholesterol levels may increase the risk of heart disease. The distribution of blood cholesterol levels in a large population of people of the same age and sex is roughly Normal. For 14-year-old boys, the mean is $\mu = 170$ milligrams of cholesterol per deciliter of blood (mg/dl) and the standard deviation is $\sigma = 30$ mg/dl. Levels above 240 mg/dl may require medical attention. What percent of 14-year-old boys have more than 240 mg/dl of cholesterol?

Solution

Call the level of cholesterol in the blood X. The variable X has a $N(170,\ 30)$ distribution. We want the proportion of boys with $X > 240$.

R code

```
1-pnorm(240,mean=170,sd=30);
```

Quantiles

Scores on the SAT verbal test in 2002 followed approximately the $N(504,\ 111)$ distribution. How high must a student score in order to place in the top 10% of all students taking the SAT?

Solution

We want to find the SAT score x^* with area 0.1 to its **right** under the Normal curve with mean $\mu = 504$ and standard deviation $\sigma = 111$. That's the same as finding the SAT score x^* with area 0.9 to its **left**.

R code

```
qnorm(0.9,mean=504,sd=111);
```